METHOD AND APPARATUS FOR SUMMARIZING MULTIPLE DOCUMENTS USING A SUBSUMPTION MODEL

FIELD OF INVENTION

The present invention relates to the field of natural language processing, information retrieval, information extraction, and automatic summary and abstraction generation.

BACKGROUND OF THE INVENTION

The advent of the Information Age has brought with it an increase in the accessibility of data, accompanied by schemes for searching that data. One searching for specific data through the Internet or in other information systems using any of many search engines available is often presented with an lengthy list of documents which may or may not contain the data for which he was searching. Reading through such a lengthy list is undesirably time consuming.

To reduce the time needlessly wasted in such reading, a variety of technologies have been presented for summarizing multiple documents to express a theme central to these documents. However, all of these technologies are inherently limited in some aspect. Some are able to search only a specific domain of knowledge and are therefore difficult to implement for different applications. Some, without radical modification, can only search documents composed in certain languages. Some use deep language parsing, statistical, or term-vector based techniques, resulting in longer waits for search results and greater demands on computing resources. Almost all generate summaries by merely

concatenating together text segments containing some keyword, often producing results which are incohesive due to anaphoric ambiguity. None use real natural language analyzing techniques. A method for summarizing multiple documents while avoiding these limitations is desirable.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention is illustrated by way of example and not limitation in the figures of the accompanying drawings, in which like references indicate similar elements, and in which:

Figure 1 is a flow diagram of one embodiment of a method for summarizing multiple documents using a subsumption model.

Figure 2 is a flow diagram of one embodiment of parsing a plurality of documents.

Figure 3 is a flow diagram of one embodiment of selecting paragraphs from the documents through subsuming relation calculation.

Figure 4 is a flow diagram of one embodiment of rewriting the selected paragraphs into a summary.

Figure 5 is an example of one embodiment of linking entity names in paragraphs of documents.

Figure 6 is an example of one embodiment of a computer system.

DETAILED DESCRIPTION

In the following description, for purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be evident, however, to one skilled in the art that the present invention may be practiced without these specific details.

Figure 1 is a flow diagram of one embodiment of a method for summarizing multiple documents using a subsumption model. In one embodiment, the content of the documents are co-related to one central topic. First, a plurality of documents are parsed, step 101. Then paragraphs are selected from the documents through subsuming relation calculation, step 102. Finally, the selected paragraphs are rewritten into a summary, step 103. Each of these steps is described in greater detail below.

Figure 2 is a flow diagram of one embodiment of parsing a plurality of documents, corresponding to step 101 of Figure 1. In one embodiment, parsing is accomplished by applying shallow natural language processing to text.

First, noun phrases and verb phrases are extracted from the documents, step 201. To accomplish this, the words in the documents are tagged according to their respective parts-of-speech. A set of rules is applied to bracket out the noun phrases and verb phrases in the documents by matching the part-of-speech tags according to predefined patterns. The noun phrases are further analyzed to identify entity names. A word with the first letter in uppercase denotes that it is part of an entity name. The use of entity name, noun phrase, and verb phrase recognition captures the features of documents while limiting the overhead involved in parsing to a minimum.

Next, the noun phrases that are entity names are categorized, step 202. Exemplary categories include people's names, company and organization names, addresses, currency amounts, dates, geographical locations, measurements, etc. In an embodiment where the documents all relate to one central topic, the detected noun phrases, verb phrases, and entity names have much in common.

Finally, the entity names are converted into canonical form, step 203. For example, "06/26/00" would be converted to "June 26, 2000". The identified entity names are input into a subsuming relation calculation.

Figure 3 is a flow diagram of one embodiment of selecting, or in other words, extracting, paragraphs from the documents through subsuming relation calculation, corresponding to step 102 of Figure 1. In one embodiment, the subsuming relation calculation is designed to calculate the inherent subsumption between paragraphs from each document. This process determines the significance of each paragraph. The noun phrases, verb phrases and/or entity names in the documents represent the content of those documents. Different paragraphs may share common noun/verb phrases and entity names. For example, if all the noun/verb phrases and entity names in a paragraph A are also in a paragraph B, then B subsumes A.

First, noun/verb phrases and entity names in each paragraph of every document are linked with identical noun/verb phrases and entity names in other paragraphs of each document, step 301. Reference links are built between the common phrases and entity names shared by paragraphs. Figure 5 discussed below illustrates an example of one embodiment of linking entity names in paragraphs of documents having a common topic independent of domain and being composed in a language other than English.

Next, the links for each paragraph are counted, step 302. The link count may be called a significance score. If a paragraph has more reference links, it is more significant than other paragraphs in representing the meaning of the documents. The more other paragraphs a given paragraph subsumes, the richer it is in content in comparison to the other paragraphs subsumed. Then, the paragraphs from the plurality of documents are ranked by their significant scores, step 303. The paragraphs with the most subsumption are relatively more dominative and informative. Therefore, these paragraphs are extracted, or in other words selected, prior to other paragraphs. In one embodiment, the top N paragraphs are bulleted, where N can be a predefined length factor decided jointly by an empirical function and a user's preference, step 304. The extracted paragraphs selected by the subsumption model are typically informative enough to represent the content of the central topic.

In one embodiment, the subsuming relation calculation is domain independent. It can process documents of a variety of topics. It does not assume any domain knowledge adaptation. Thus, it is relatively easy to implement for different applications.

Unlike other summarization systems, no statistic technique is used in the subsuming relation calculation. Therefore, no background corpus is needed to build a base frequency. The domain and length of the documents are not limited. The subsuming relation calculation is also not term-vector based, avoiding high dimension vector manipulation.

Figure 4 is a flow diagram of one embodiment of rewriting the selected paragraphs into a summary, corresponding to step 103 of Figure 1. First, the paragraphs are ranked, step 401, by their significance score. In one embodiment, the top N

paragraphs are bulleted, where N can be a predefined length factor decided jointly by an empirical function and a user's preference. Cohesiveness is less likely if these bulleted paragraphs are output as a summary without further processing. So a co-reference resolution algorithm is applied to the paragraphs, step 402, to resolve anaphoric ambiguity. There are a number of such algorithms in the public domain. By introducing the co-reference resolution, most anaphoric ambiguity is removed, thus making the result summary more cohesive.

For example, a document might read, "I met John and Mary this morning. He was driving a red car. It's a nice sports car. She was very happy." A reader may not notice any co-reference ambiguity in it, since it's obviously that "*he*" stands for *John*, "*she*" stands for *Mary* and "*it*" stands for the car. But the method and apparatus disclosed herein extracts the significant paragraphs (or sentences) for multiple documents and concatenates them into one text passage as a summary, and because these paragraphs may come from different documents, or different parts of the same document, they may contain pronouns that may refer to entity names in paragraphs that were not extracted and do not appear in the resulting summary. To reduce reader confusion, a one-to-one reference relation is built between each pronoun and its equivalent entity name.

Finally, the pronouns (for example, he, she, it, they, etc.) in the paragraphs are replaced with their full entity name antecedents, step 403. Thus, the readability of the output summary is improved.

The subsuming relation calculation can be applied to languages other than English. To apply the calculation to another language, only the shallow natural language

processing and co-reference resolution components need to be modified. The core subsumption model is language independent.

Figure 5 is an example of one embodiment of linking identical entity names in paragraphs of documents having a common topic independent of domain and being composed in a language other than English. One paragraph 501 contains entity names which are also contained in another paragraph 502. The identical entity names in each paragraph are linked according to the flow diagram in Figure 3. Because all of the entity names in paragraph 501 are also contained in paragraph 502, paragraph 502 can be said to subsume paragraph 501.

The method and apparatus disclosed herein may be integrated into advanced Internet- or network-based knowledge systems as related to information retrieval, information extraction, and question and answer systems. Figure 6 is an example of one embodiment of a computer system. The system shown has a processor 601 coupled to a bus 602. Also shown coupled to the bus are a memory 603 which may contain instructions 604. Additional components shown coupled to the bus are a storage device 605 (such as a hard drive, floppy drive, CD-ROM, DVD-ROM, etc.), an input device 606 (such as a keyboard, mouse, light pen, bar code reader, scanner, microphone, joystick, etc.), and an output device (such as a printer, monitor, speakers, etc.). Of course, an exemplary computer system could have more components than these or a subset of the components listed.

The method described above can be stored in the memory of a computer system (e.g., set top box, video recorders, etc.) as a set of instructions to be executed, as shown by way of example in Figure 6. In addition, the instructions to perform the method

described above could alternatively be stored on other forms of machine-readable media, including magnetic and optical disks. For example, the method of the present invention could be stored on machine-readable media, such as magnetic disks or optical disks, which are accessible via a disk drive (or computer-readable medium drive). Further, the instructions can be downloaded into a computing device over a data network in a form of compiled and linked version.

Alternatively, the logic to perform the methods as discussed above, could be implemented in additional computer and/or machine readable media, such as discrete hardware components as large-scale integrated circuits (LSI's), application-specific integrated circuits (ASIC's), firmware such as electrically erasable programmable read-only memory (EEPROM's); and electrical, optical, acoustical and other forms of propagated signals (e.g., carrier waves, infrared signals, digital signals, etc.); etc.

Although the present invention has been described with reference to specific exemplary embodiments, it will be evident that various modifications and changes may be made to these embodiments without departing from the broader spirit and scope of the invention. Accordingly, the specification and drawings are to be regarded in an illustrative rather than a restrictive sense.